Fast Nonsmooth Regularized Risk Minimization with Continuation Appendix

Shuai Zheng Ruiliang Zhang James T. Kwok
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong
{szhengac, rzhangaf, jamesk}@cse.ust.hk

Note that, for a batch solver, $\kappa_s = L_s/\mu$ for μ-strongly convex objectives (where $L_s = \hat{L} + \frac{\|A\|_2^2}{\zeta\gamma_s}$ is the Lipschitz constant of $\tilde{f}_{\gamma_s}$). When the solver is stochastic, we use $\kappa_s = L_{m,s}/\mu$, where $L_{m,s} = \max_i \hat{L}_i + \frac{\|A_i\|_2^2}{\zeta\gamma_s}$ (Schmidt, Roux, and Bach 2013).

Proof of Lemma 1

Lemma 1. *For both non-accelerated solvers and accelerated solvers, if T_1 is large enough such that $\rho_1 \leq \tilde{\rho}$, where $\tilde{\rho} \in (0,1)$, then $\rho_s \leq \tilde{\rho}$ for all $s > 1$.*

Proof. To prove this result, we use induction.

- Non-accelerated solvers:
  Base step: Since we assume that T_1 is large enough such that $\rho_1 \leq \tilde{\rho}$, then property holds for $s = 1$.
  Inductive step: Assume that $\rho_s \leq \tilde{\rho}$. Consider ρ_{s+1}. By the definition of κ_s and $\gamma_{s+1} = \gamma_s/\tau$, we have $\kappa_{s+1} \leq \tau\kappa_s$. Recall that the non-accelerated solvers take $T_s = a\kappa_s\phi(\rho_s) + b\phi(\rho_s) + c$ iterations to achieve error reduction factor ρ_s at stage s. With the assumptions on $\phi(\rho_s)$ and a, b, c, we have $\phi(\rho_s) \geq \phi(\tilde{\rho})$ and $T_s \geq a\kappa_s\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c$. By $T_{s+1} = \tau T_s$, we have

$$\begin{aligned}
T_{s+1} &= a\kappa_{s+1}\phi(\rho_{s+1}) + b\phi(\rho_{s+1}) + c \\
&= a\tau\kappa_s\phi(\rho_s) + b\tau\phi(\rho_s) + c\tau \\
&\geq a\tau\kappa_s\phi(\tilde{\rho}) + b\tau\phi(\tilde{\rho}) + c\tau \\
&\geq a\kappa_{s+1}\phi(\tilde{\rho}) + b\tau\phi(\tilde{\rho}) + c\tau \\
&\geq a\kappa_{s+1}\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c
\end{aligned}$$

  Hence, we have $\rho_{s+1} \leq \tilde{\rho}$.
- Accelerated solvers: The first part of the proof is identical to that for non-accelerated solvers. By $T_{s+1} = \sqrt{\tau}T_s$, we have

$$\begin{aligned}
T_{s+1} &= a\sqrt{\kappa_{s+1}}\phi(\rho_{s+1}) + b\phi(\rho_{s+1}) + c \\
&= a\sqrt{\tau\kappa_s}\phi(\rho_s) + b\sqrt{\tau}\phi(\rho_s) + c\sqrt{\tau} \\
&\geq a\sqrt{\tau\kappa_s}\phi(\tilde{\rho}) + b\sqrt{\tau}\phi(\tilde{\rho}) + c\sqrt{\tau} \\
&\geq a\sqrt{\kappa_{s+1}}\phi(\tilde{\rho}) + b\sqrt{\tau}\phi(\tilde{\rho}) + c\sqrt{\tau} \\
&\geq a\sqrt{\kappa_{s+1}}\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c
\end{aligned}$$

  Hence, we have $\rho_{s+1} \leq \tilde{\rho}$.

□

Proof of Lemma 2

Lemma 2. $\tilde{P}_s(x) - \tilde{P}_s(x_s^*) - \gamma_s D_u \leq P(x) - P(x^*) \leq \tilde{P}_s(x) - \tilde{P}_s(x_s^*) + \gamma_s D_u$.

Proof. Following immediately from (2.7) of (Nesterov 2005), we have $P(x) - P(x^*) \leq \tilde{P}_s(x) - \tilde{P}_s(x_s^*) + \gamma_s D_u$ for any $x \in \mathbb{R}^d$.

From (Nesterov 2005), $P(x) \leq \tilde{P}_s(x) + \gamma_s D_u$. Thus, $-\tilde{P}_s(x_s^*) \leq -P(x_s^*) + \gamma_s D_u \leq -P(x^*) + \gamma_s D_u$. Combining with the fact in (Nesterov 2005) that $\tilde{P}_s(x) \leq P(x)$, we have $\tilde{P}_s(x) - \tilde{P}_s(x_s^*) \leq P(x) - P(x^*) + \gamma_s D_u$.

Thus, we have $\tilde{P}_s(x) - \tilde{P}_s(x_s^*) - \gamma_s D_u \leq P(x) - P(x^*) \leq \tilde{P}_s(x) - \tilde{P}_s(x_s^*) + \gamma_s D_u$

□

Proof of Theorem 1

Lemma 3. *If γ_s is monotonically decreasing with s, then for any $s \geq 2$ and $x \in \mathbb{R}^d$,*

$$\tilde{P}_s(x) - \tilde{P}_s(x_s^*) \leq \tilde{P}_{s-1}(x) - \tilde{P}_{s-1}(x_{s-1}^*) + (\gamma_{s-1} - \gamma_s)D_u.$$

Proof. From Lemma 9 in (Ouyang and Gray 2012), we have $\tilde{P}_{s-1}(x) \leq \tilde{P}_s(x) \leq \tilde{P}_{s-1}(x) + (\gamma_{s-1} - \gamma_s)D_u$. Result follows by combining the two parts of the inequality. □

Proof. (of Theorem 1) With $\gamma_s = \frac{\gamma_1}{\tau^{s-1}}$, we have

$$\begin{aligned}
&\mathbb{E}P(\tilde{x}_s) - P(x^*) \\
\leq\;& \mathbb{E}\tilde{P}_s(\tilde{x}_s) - \tilde{P}_s(x_s^*) + \gamma_s D_u \quad \text{(by Lemma 2)} \\
\leq\;& \rho_s(\mathbb{E}\tilde{P}_s(\tilde{x}_{s-1}) - \tilde{P}_s(x_s^*)) + \gamma_s D_u \quad \text{(by Assumption 2)} \\
\leq\;& \rho_s(\mathbb{E}\tilde{P}_{s-1}(\tilde{x}_{s-1}) - \tilde{P}_{s-1}(x_{s-1}^*) + (\gamma_{s-1} - \gamma_s)D_u) + \gamma_s D_u \quad \text{(by Lemma 3)} \\
=\;& \rho_s(\mathbb{E}\tilde{P}_{s-1}(\tilde{x}_{s-1}) - \tilde{P}_{s-1}(x_{s-1}^*)) + \rho_s(\gamma_{s-1} - \gamma_s)D_u + \gamma_s D_u \\
\leq\;& \rho_s\rho_{s-1}(\mathbb{E}\tilde{P}_{s-1}(\tilde{x}_{s-2}) - \tilde{P}_{s-1}(x_{s-1}^*)) + \rho_s(\gamma_{s-1} - \gamma_s)D_u + \gamma_s D_u \quad \text{(by Assumption 2)} \\
\leq\;& \left(\prod_{i=1}^{s}\rho_i\right)(\tilde{P}_1(\tilde{x}_0) - \tilde{P}_1(x_1^*)) + \left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}}\right)\gamma_1 D_u \\
\leq\;& \left(\prod_{i=1}^{s}\rho_i\right)(P(\tilde{x}_0) - P(x^*)) + \underbrace{\left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}} + \prod_{i=1}^{s}\rho_i\right)}_{\beta_s}\gamma_1 D_u, \qquad (1)
\end{aligned}$$

where in the second-to-last inequality, we apply Lemma 3 and Assumption 2 recursively the same as second and third inequality, and use $\gamma_{s-1} - \gamma_s = \frac{\tau-1}{\tau^{s-1}}\gamma_1$. In the last inequality, we use Lemma 2. Moreover, note that $\{\beta_s\}$ is monotonically decreasing as follows.

$$\begin{aligned}
\beta_s - \beta_{s-1} &= \left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}} + \prod_{i=1}^{s}\rho_i\right) - \left(\sum_{i=1}^{s-2}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s-1}\rho_j + \frac{1}{\tau^{s-2}} + \prod_{i=1}^{s-1}\rho_i\right) \\
&= \left(\sum_{i=1}^{s-2}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s-1}\rho_j\right)(\rho_s - 1) + \frac{(\tau-1)(\rho_s-1)}{\tau^{s-1}} + \left(\prod_{i=1}^{s-1}\rho_i\right)(\rho_s - 1) \\
&< 0.
\end{aligned}$$

Hence, from (1), $\mathbb{E}P(\tilde{x}_s) - P(x^*)$ converges to zero. We now find out how fast $\{\beta_s\}$ decays. Let $\tilde{\rho} = \frac{1}{\tau^2}$, we obtain

$$\begin{aligned}
\beta_s &= \sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}} + \prod_{i=1}^{s}\rho_i \\
&\leq \sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\tilde{\rho}^{s-i} + \frac{1}{\tau^{s-1}} + \tilde{\rho}^s \quad \text{(by Lemma 1)} \qquad (2) \\
&= \sum_{i=1}^{s-1}\frac{\tau-1}{\tau^{2s-i}} + \frac{1}{\tau^{s-1}} + \frac{1}{\tau^{2s}} \\
&= \frac{1}{\tau^s} - \frac{1}{\tau^{2s-1}} + \frac{1}{\tau^{s-1}} + \frac{1}{\tau^{2s}} \\
&\leq \frac{1+\tau}{\tau^s}, \qquad (3)
\end{aligned}$$

and

$$T = \sum_{i=1}^{s} T_i = T_1\sum_{i=1}^{s}\tau^{i-1} = \frac{\tau^s - 1}{\tau - 1}T_1. \qquad (4)$$

These imply $s = O(\log(T))$ and $\beta_s = O\left(\frac{1}{T}\right)$. From (1), we obtain

$$\mathbb{E}P(\tilde{x}_s) - P(x^*) \quad \leq \quad \left(\prod_{i=1}^{s} \rho_i\right)(P(\tilde{x}_0) - P(x^*)) + O\left(\frac{\gamma_1 D_u}{T}\right).$$

□

Proof of Theorem 2

Proof. The first part of the proof is identical to that for Theorem 1. Here, as $T_s = \sqrt{\tau} T_{s-1}$, we have

$$T = \sum_{i=1}^{s} T_i = T_1 \sum_{i=1}^{s} \sqrt{\tau}^{i-1} = \frac{\sqrt{\tau}^s - 1}{\sqrt{\tau} - 1} T_1. \tag{5}$$

Hence, $s = O(\log(T))$, $\beta_s = O\left(\frac{1}{T^2}\right)$, and (1) yeilds

$$\mathbb{E}P(\tilde{x}_s) - P(x^*) \leq \left(\prod_{i=1}^{s} \rho_i\right)(P(\tilde{x}_0) - P(x^*)) + O\left(\frac{\gamma_1 D_u}{T^2}\right).$$

□

Proposition 6. *If we require $\rho_1 \leq 1/\tau$, the rate will be slowed to $O(\log T/T)$; if $\rho_1 \leq 1/\sqrt{\tau}$, it degrades further to $O(1/\sqrt{T})$. On the other hand, if $\rho_1 \leq 1/\tau^c$ with $c > 2$, the rate remains at $O(1/T)$.*

Proof. Following (2) and (4),

- if $\tilde{\rho} = \frac{1}{\tau}$, then it leads to $\beta_s \leq \frac{s(\tau-1)+2}{\tau^s} = O(\log T/T)$.
- if $\tilde{\rho} = \frac{1}{\sqrt{\tau}}$, then $\beta_s \leq \frac{\sqrt{\tau}+2}{\sqrt{\tau}^s} = O(1/\sqrt{T})$.
- if $\tilde{\rho} = \frac{1}{\tau^c}$ with $c > 2$, then $\beta_s \leq \frac{\tau+1}{\tau^s} = O(1/T)$.

□

Proposition 7. *If we require $\rho_1 \leq 1/\tau$, the rate will be slowed to $O(\log T/T^2)$; if $\rho_1 \leq 1/\sqrt{\tau}$, it degrades further to $O(1/T)$. On the other hand, if $\rho_1 \leq 1/\tau^c$ with $c > 2$, the rate remains at $O(1/T^2)$.*

Proof. Following (2) and (5),

- if $\tilde{\rho} = \frac{1}{\tau}$, then it leads to $\beta_s \leq \frac{s(\tau-1)+2}{\tau^s} = O(\log T/T^2)$.
- if $\tilde{\rho} = \frac{1}{\sqrt{\tau}}$, then $\beta_s \leq \frac{\sqrt{\tau}+2}{\sqrt{\tau}^s} = O(1/T)$.
- if $\tilde{\rho} = \frac{1}{\tau^c}$ with $c > 2$, then $\beta_s \leq \frac{\tau+1}{\tau^s} = O(1/T^2)$.

□

Proof of Theorem 3

In this section, x_s^* denotes the optimal solution to $H_s(x)$.

Note that there are two cases regarding condtion number κ_s. If $\frac{\lambda_s}{2}\|x\|_2^2$ is added to $\tilde{f}_{\gamma_s}$, $\kappa_s = (L_s + \lambda_s)/\lambda_s$ for batch solvers and $\kappa_s = (L_{m,s} + \lambda_s)/\lambda_s$ for stochastic solvers, or if $\frac{\lambda_s}{2}\|x\|_2^2$ is added to r, $\kappa_s = L_s/\lambda_s$ for batch solvers and $\kappa_s = L_{m,s}/\lambda_s$ for stochastic solvers.

Lemma 4. *For any $x \in \mathbb{R}^d$,*

$$P(x) - P(x^*) \leq H_s(x) - H_s(x_s^*) + \gamma_s D_u + \frac{\lambda_s}{2}\|x^*\|_2^2,$$

Proof. As $\tilde{P}_s(x) \leq P(x) \leq \tilde{P}_s(x) + \gamma_s D_u$ by (2.7) of (Nesterov 2005), we have $P(x) \leq H_s(x) + \gamma_s D_u$, and also $H_s(x_s^*) = \tilde{P}_s(x_s^*) + \frac{\lambda_s}{2}\|x_s^*\|_2^2 \leq \min_x P(x) + \frac{\lambda_s}{2}\|x\|_2^2 \leq P(x^*) + \frac{\lambda_s}{2}\|x^*\|_2^2$. Result follows on combining the two inequalities. □

Lemma 5. *For any $x \in \mathbb{R}^d$, $H_s(x) - H_s(x_s^*) \leq P(x) - P(x^*) + \gamma_s D_u + \frac{\lambda_s}{2}\|x\|_2^2$.*

Proof. Since $\tilde{P}_s(x) \leq P(x)$, we have $H_s(x) \leq P(x) + \frac{\lambda_s}{2}\|x\|_2^2$. Moreover, since $P(x) \leq H_s(x) + \gamma_s D_u$, and so $P(x^*) \leq H_s(x_s^*) + \gamma_s D_u$. Result follows on combining the two inequalities. □

Lemma 6. *If γ_s and λ_s are monotonically decreasing with s, then for any $s \geq 2$ and $x \in \mathbb{R}^d$,*

$$H_s(x) - H_s(x_s^*) \leq H_{s-1}(x) - H_{s-1}(x_{s-1}^*) + (\gamma_{s-1} - \gamma_s)D_u + \frac{1}{2}(\lambda_{s-1} - \lambda_s)\|x_s^*\|^2,$$

Proof. From Lemma 9 in (Ouyang and Gray 2012), we have $\tilde{P}_{s-1}(x) \leq \tilde{P}_s(x) \leq \tilde{P}_{s-1}(x) + (\gamma_{s-1} - \gamma_s)D_u$. Since $\lambda_{s-1} > \lambda_s$, then

$$H_s(x) \leq H_{s-1}(x) + (\gamma_{s-1} - \gamma_s)D_u.$$

Moreover, $\tilde{P}_{s-1}(x) \leq \tilde{P}_s(x)$ implies $H_{s-1}(x) + \frac{1}{2}(\lambda_s - \lambda_{s-1})\|x\|^2 \leq H_s(x)$. Thus,

$$H_{s-1}(x_{s-1}^*) \leq H_s(x_s^*) + \frac{1}{2}(\lambda_{s-1} - \lambda_s)\|x_s^*\|^2.$$

Result follows on combining the two inequalities. □

Lemma 7. *For both non-accelerated solvers and accelerated solvers, if T_1 is large enough such that $\rho_1 \leq \tilde{\rho}$, where $\tilde{\rho} \in (0, 1)$, then $\rho_s \leq \tilde{\rho}$ for all $s > 1$.*

Proof. The proof is similar to the one of Lemma 1. We consider induction.

- Non-accelerated solvers:

  Base step: Since we assume that T_1 is large enough such that $\rho_1 \leq \tilde{\rho}$, then property holds for $s = 1$.

  Inductive step: Assume that $\rho_s \leq \tilde{\rho}$. Consider ρ_{s+1}. By the definition of κ_s, $\gamma_{s+1} = \gamma_s/\tau$ and $\lambda_{s+1} = \lambda_s/\tau$, we have $\kappa_{s+1} \leq \tau^2\kappa_s$. Recall that the non-accelerated solvers take $T_s = a\kappa_s\phi(\rho_s) + b\phi(\rho_s) + c$ iterations to achieve error reduction factor ρ_s at stage s. With the assumptions on $\phi(\rho_s)$ and a, b, c, we have $\phi(\rho_s) \geq \phi(\tilde{\rho})$ and $T_s \geq a\kappa_s\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c$. By $T_{s+1} = \tau^2 T_s$, we have

$$\begin{aligned} T_{s+1} &= a\kappa_{s+1}\phi(\rho_{s+1}) + b\phi(\rho_{s+1}) + c \\ &= a\tau^2\kappa_s\phi(\rho_s) + b\tau^2\phi(\rho_s) + c\tau^2 \\ &\geq a\tau^2\kappa_s\phi(\tilde{\rho}) + b\tau^2\phi(\tilde{\rho}) + c\tau^2 \\ &\geq a\kappa_{s+1}\phi(\tilde{\rho}) + b\tau^2\phi(\tilde{\rho}) + c\tau^2 \\ &\geq a\kappa_{s+1}\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c \end{aligned}$$

  Hence, we have $\rho_{s+1} \leq \tilde{\rho}$.

- Accelerated solvers: The first part of the proof is identical to that for non-accelerated solvers. By $T_{s+1} = \tau T_s$, we have

$$\begin{aligned} T_{s+1} &= a\sqrt{\kappa_{s+1}}\phi(\rho_{s+1}) + b\phi(\rho_{s+1}) + c \\ &= a\sqrt{\tau^2\kappa_s}\phi(\rho_s) + b\tau\phi(\rho_s) + c\tau \\ &\geq a\sqrt{\tau^2\kappa_s}\phi(\tilde{\rho}) + b\tau\phi(\tilde{\rho}) + c\tau \\ &\geq a\sqrt{\kappa_{s+1}}\phi(\tilde{\rho}) + b\tau\phi(\tilde{\rho}) + c\tau \\ &\geq a\sqrt{\kappa_{s+1}}\phi(\tilde{\rho}) + b\phi(\tilde{\rho}) + c \end{aligned}$$

  Hence, we have $\rho_{s+1} \leq \tilde{\rho}$.

□

Proof. (of Theorem 3) With $\gamma_s = \frac{\gamma_1}{\tau^{s-1}}$, $\lambda_s = \frac{\lambda_1}{\tau^{s-1}}$, we have

$$\begin{aligned}
&\mathbb{E}P(\tilde{x}_s) - P(x^*) \\
\leq\ & \mathbb{E}H_s(\tilde{x}_s) - H_s(x_s^*) + \gamma_s D_u + \frac{\lambda_s}{2}\|x^*\|_2^2 \quad \text{(by Lemma 4)} \\
\leq\ & \rho_s\left(\mathbb{E}H_s(\tilde{x}_{s-1}) - H_s(x_s^*)\right) + \gamma_s D_u + \frac{\lambda_s}{2}\|x^*\|_2^2 \quad \text{(by Assumption 3)} \\
\leq\ & \rho_s\left(\mathbb{E}H_{s-1}(\tilde{x}_{s-1}) - H_{s-1}(x_{s-1}^*) + (\gamma_{s-1}-\gamma_s)D_u + (\lambda_{s-1}-\lambda_s)\frac{1}{2}\|x_s^*\|_2^2\right) \\
& + \gamma_s D_u + \frac{\lambda_s}{2}\|x^*\|_2^2 \quad \text{(by Lemma 6)} \\
=\ & \rho_s\left(\mathbb{E}H_{s-1}(\tilde{x}_{s-1}) - H_{s-1}(x_{s-1}^*)\right) + \rho_s(\gamma_{s-1}-\gamma_s)D_u + \gamma_s D_u \\
& + \rho_s(\lambda_{s-1}-\lambda_s)\frac{1}{2}\|x_s^*\|_2^2 + \frac{\lambda_s}{2}\|x^*\|_2^2 \\
\leq\ & \rho_s\rho_{s-1}\left(\mathbb{E}H_{s-1}(\tilde{x}_{s-2}) - H_{s-1}(x_{s-1}^*)\right) + \rho_s(\gamma_{s-1}-\gamma_s)D_u + \gamma_s D_u \\
& + \rho_s(\lambda_{s-1}-\lambda_s)\frac{1}{2}\|x_s^*\|_2^2 + \frac{\lambda_s}{2}\|x^*\|_2^2 \quad \text{(by Assumption 3)} \\
\leq\ & \left(\prod_{i=1}^{s}\rho_i\right)\left(H_1(\tilde{x}_0) - H_1(x_1^*)\right) + \left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}}\right)\gamma_1 D_u \\
& + \left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}}\right)\frac{\lambda_1}{2}R^2 \\
\leq\ & \left(\prod_{i=1}^{s}\rho_i\right)\left(P(\tilde{x}_0) - P(x^*)\right) + \left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}} + \prod_{i=1}^{s}\rho_i\right)\gamma_1 D_u \\
& + \left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}}\right)\frac{\lambda_1}{2}R^2 + \left(\prod_{i=1}^{s}\rho_i\right)\frac{\lambda_1}{2}\|\tilde{x}_0\|_2^2 \\
=\ & \left(\prod_{i=1}^{s}\rho_i\right)\left(P(\tilde{x}_0) - P(x^*) + \frac{\lambda_1}{2}\|\tilde{x}_0\|_2^2\right) + \underbrace{\left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}} + \prod_{i=1}^{s}\rho_i\right)}_{\beta_s}\gamma_1 D_u \\
& + \underbrace{\left(\sum_{i=1}^{s-1}\frac{\tau-1}{\tau^i}\prod_{j=i+1}^{s}\rho_j + \frac{1}{\tau^{s-1}}\right)}_{\alpha_s}\frac{\lambda_1}{2}R^2, \qquad (6)
\end{aligned}$$

where in the second-to-last inequality, we apply Lemma 6 and Assumption 3 recursively the same as second and third inequality, and use $\gamma_{s-1} - \gamma_s = \frac{\tau-1}{\tau^{s-1}}\gamma_1$ and $\lambda_{s-1} - \lambda_s = \frac{\tau-1}{\tau^{s-1}}\lambda_1$, and apply assumption $\|x^*\|_2 \leq R$ and $\|x_s^*\|_s \leq R$ for all s. In the last inequality, we use Lemma 5. By the proof of Theorem 1 and Lemma 7 with $\tilde{\rho} = \frac{1}{\tau^2}$, we have $\beta_s, \alpha_s \leq \frac{1+\tau}{\tau^s}$. And

$$T = \sum_{i=1}^{s} T_i = T_1 \sum_{i=1}^{s} \tau^{2(i-1)} = \frac{\tau^{2s}-1}{\tau^2-1}T_1, \qquad (7)$$

which implys that $s = O(\log(T))$ and $\beta_s, \alpha_s = O\left(\frac{1}{\sqrt{T}}\right)$. Then, we obtain

$$\mathbb{E}P(\tilde{x}_s) - P(x^*) \leq \left(\prod_{i=1}^{s}\rho_i\right)\left(P(\tilde{x}_0) - P(x^*) + \frac{\lambda_1}{2}\|\tilde{x}_0\|_2^2\right) + O\left(\frac{\lambda_1 R^2}{\sqrt{T}}\right) + O\left(\frac{\gamma_1 D_u}{\sqrt{T}}\right).$$

For the convergence rate of accelerated solvers, the first part of the proof is identical to that for non-accelerated solvers. Here, as $T_s = \tau T_{s-1}$, we have

$$T = \sum_{i=1}^{s} T_i = T_1 \sum_{i=1}^{s} \tau^{i-1} = \frac{\tau^s - 1}{\tau - 1} T_1 \tag{8}$$

Hence, $s = O(\log(T))$, $\beta_s, \alpha_s = O\left(\frac{1}{T}\right)$, and (6) yields

$$\mathbb{E}P(\tilde{x}_s) - P(x^*) \leq \left(\prod_{i=1}^{s} \rho_i\right)\left(P(\tilde{x}_0) - P(x^*) + \frac{\lambda_1}{2}\|\tilde{x}_0\|_2^2\right) + O\left(\frac{\lambda_1 R^2}{T}\right) + O\left(\frac{\gamma_1 D_u}{T}\right).$$

□

Proposition 8. *For non-accelerated solvers, If we require $\rho_1 \leq 1/\tau$, the rate will be slowed to $O(\log T/\sqrt{T})$; if $\rho_1 \leq 1/\sqrt{\tau}$, it degrades further to $O(1/T^{1/4})$. On the other hand, if $\rho_1 \leq 1/\tau^c$ with $c > 2$, the rate remains at $O(1/\sqrt{T})$.*

For accelerated solvers, If we require $\rho_1 \leq 1/\tau$, the rate will be slowed to $O(\log T/T)$; if $\rho_1 \leq 1/\sqrt{\tau}$, it degrades further to $O(1/\sqrt{T})$. On the other hand, if $\rho_1 \leq 1/\tau^c$ with $c > 2$, the rate remains at $O(1/T)$.

Proof. Following the proof of Proposition 6 and 7 with (7) and (8). □

Convergence Factors of Example Algorithms

- Proximal Gradient descent (Nesterov 2013): $O(\kappa_s \phi(\rho_s)) = 4\kappa_s \log(1/\rho_s)$
- Accelerated Proximal Gradient descent (Nesterov 2004; Schmidt, Roux, and Bach 2011): $O(\sqrt{\kappa_s}\phi(\rho_s)) = \sqrt{\kappa_s}\log(2/\rho_s)$
- Proximal SVRG (Xiao and Zhang 2014): $O(\kappa_s \phi(\rho_s)) = \frac{\theta}{(1-4\theta)\rho_s - 4\theta}(\kappa_s + 4)$
- Accelerated Proximal SVRG (Nitanda 2014): $O(\kappa_s \phi(\rho_s)) = \sqrt{\kappa_s}\frac{\sqrt{2}}{(1-p)}\log(\frac{1}{\frac{\rho_s}{2+p} - \frac{p}{1-p}})$
- SAGA (Defazio, Bach, and Lacoste-Julien 2014): $O(\kappa_s \phi(\rho_s)) = \frac{3n}{\rho_s}\left(\frac{3\kappa_s}{n} + 1\right)$
- MISO (Mairal 2013): $O(\kappa_s \phi(\rho_s)) = \frac{n\kappa_s}{\rho_s}$

where $\theta \in (0, 0.25)$ and satisfies $(1-4\theta)\rho_s - 4\theta > 0$, and $p \in (0,1)$ and satisfies $\rho_s > \frac{p(2+p)}{1-p}$. The convergence rate for SAGA and MISO on strongly convex problems are derived from each convergence rate on general convex problems with some mathematical transformations.

For SAGA:

$$\begin{aligned}
&\mathbb{E}\tilde{P}_s(\tilde{x}_s) - \tilde{P}_s(x_s^*) \\
&\leq \frac{3n}{T_s}\left[\frac{3L_{m,s}}{2n}\|\tilde{x}_{s-1} - x_s^*\|_2^2 + \tilde{f}_{\gamma_s}(\tilde{x}_{s-1}) - \nabla \tilde{f}_{\gamma_s}(x_s^*)^T(\tilde{x}_{s-1} - x_s^*) - \tilde{f}_{\gamma_s}(x_s^*)\right] \quad \text{(by (Defazio, Bach, and Lacoste-Julien 2014))} \\
&\leq \frac{3n}{T_s}(\frac{3L_{m,s}}{n\mu} + 1)\left(\tilde{P}_s(\tilde{x}_{s-1}) - \tilde{P}_s(x_s^*)\right)
\end{aligned}$$

where second inequality come from $\frac{\mu}{2}\|\tilde{x}_{s-1} - x_s^*\|_2^2 \leq \tilde{P}_s(\tilde{x}_{s-1}) - \tilde{P}_s(x_s^*)$ and $-\nabla \tilde{f}_{\gamma_s}(x_s^*)^T(\tilde{x}_{s-1} - x_s^*) \leq r(\tilde{x}_{s-1}) - r(x_s^*)$.

For MISO:

$$\begin{aligned}
&\mathbb{E}\tilde{P}_s(\tilde{x}_s) - \tilde{P}_s(x_s^*) \\
&\leq \frac{nL_{m,s}}{2T_s}\|\tilde{x}_{s-1} - x_s^*\|_2^2 \quad \text{(by (Mairal 2013))} \\
&\leq \frac{nL_{m,s}}{T_s\mu}\left(\tilde{P}_s(\tilde{x}_{s-1}) - \tilde{P}_s(x_s^*)\right)
\end{aligned}$$

References

Defazio, A.; Bach, F.; and Lacoste-Julien, S. 2014. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2116–2124.

Mairal, J. 2013. Optimization with first-order surrogate functions. In *Proceedings of the 30th International Conference on Machine Learning*.

Nesterov, Y. 2004. *Introductory Lectures on Convex Optimization*, volume 87. Springer.

Nesterov, Y. 2005. Smooth minimization of non-smooth functions. *Mathematical Programming* 103(1):127–152.

Nesterov, Y. 2013. Gradient methods for minimizing composite functions. *Mathematical Programming* 140(1):125–161.

Nitanda, A. 2014. Stochastic proximal gradient descent with acceleration techniques. In *Advances in Neural Information Processing Systems*, 1574–1582.

Ouyang, H., and Gray, A. 2012. Stochastic smoothing for nonsmooth minimizations: Accelerating SGD by exploiting structure. *arXiv preprint arXiv:1205.4481*.

Schmidt, M.; Roux, N. L.; and Bach, F. R. 2011. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, 1458–1466.

Schmidt, M.; Roux, N. L.; and Bach, F. 2013. Minimizing finite sums with the stochastic average gradient. arXiv Preprint arXiv:1309.2388.

Xiao, L., and Zhang, T. 2014. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization* 24(4):2057–2075.